THE REVISED OSTERLIND INDEX. A COMPARATIVE ANALYSIS IN CONTENT VALIDITY STUDIES[1]

EL ÍNDICE DE OSTERLIND REVISADO. UN ANÁLISIS COMPARATIVO EN ESTUDIO DE VALIDEZ DE CONTENIDO

Susana Sanduvete-Chaves, Salvador Chacón-Moscoso, Milagrosa Sánchez-Martín y José Antonio Pérez-Gil (†)

Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla

Cómo referenciar este artículo/How to reference this article:

Sanduvete-Chaves, S., Chacón-Moscoso, S., Sánchez-Martín, M. y Pérez-Gil, J. A. (2013). The Revised Osterlind Index. A Comparative Analysis in Content Validity Studies [El Indice de Osterlind Revisado. Un Análisis Comparativo en Estudios de Validez de Contenido]. *Acción Psicológica, 10*(2), 19-26. http://dx.doi.org/10.5944/ap.10.2.11821

Abstract

A procedure commonly used to obtain empirical evidence in content validity studies is the calculation of the Osterlind index after gathering the expert opinions about the adequacy of the items to measure a particular dimension of the tool. The aim of this work is to compare the results obtained when experts score the degree of suitability item-dimension on the traditional 3-point rating scales with the results obtained using, alternatively, 5-point ones. 105 participants valued, on 5-point rating scales, the fitness item-dimension of 31 items to 7 dimensions that composed a questionnaire to measure satisfaction with the training received. These marks were also transformed into 3-point rating scales. Comparison between Osterlind indexes calculated using scores from 5 and 3-point rating scales shows that the new propose is more conservative than the classic procedure; i.e., items considered adequate using 3-point rating scales were removed using 5-point rating scales.

Keywords: Content Validity, Osterlind Index, Rating Scales, Empirical Evidence, Comparison.

Resumen

Un procedimiento comúnmente usado para obtener evidencias empíricas en estudios de validez de contenido es el cálculo del índice de Osterlind tras recoger la opinión de expertos acerca de la adecuación de ítems para medir una concreta dimensión de un test. El objetivo de este trabajo es comparar los resultados obtenidos cuando los expertos puntúan el grado

Correspondencia: Susana Sanduvete-Chaves. Departamento de Psicología Experimental, Facultad de Psicología, Universidad de Sevilla. Email: sussancha@us.es.

Recibido: 12/05/2013
Aceptado: 04/07/2013

1 This study forms part of the results obtained in research project PSI2011-29587, funded by Spain's Ministry for Science and Innovation.

de adecuación ítem-dimensión sobre la tradicional escala de valoración de 3 puntos y sobre una de 5 puntos. 105 participantes valoraron, sobre escalas de valoración de 5 puntos, la adecuación de 31 ítems a 7 dimensiones de un cuestionario para medir la satisfacción con la formación recibida. Estas puntuaciones fueron posteriormente transformadas a 3 puntos de valoración. La comparación entre los índices de Osterlind calculados a partir de valoraciones sobre escalas de 5 y 3 puntos muestra que, usando las escalas de 5 puntos, se eliminaron ítems que fueron considerados adecuados usando escalas de 3.

Palabras Clave: Validez de Contenido, Índice de Osterlind, Escalas de Valoración, Evidencias Empíricas, Comparación.

Introduction

The need of testing validity of psychological instruments is commonly accepted (Carrasco, Holgado, Del Barrio, & Barbero, 2008). The term validity refers to the approximate truth of an inference (Shadish, Cook, & Campbell, 2002); concretely, content validity can be defined as the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured (Anastasi & Urbina, 1997).

The adequacy of the items that compose a test to content validity criteria implies (Chacón, Pérez-Gil, Holgado, & Lara, 1991): (a) The development of those items is based on the theoretical framework that supports the concept of measuring (that is why the concept must be defined in a clear and precise way); and (b) In an operational way, the synthesis of opinions by a group of experts (researchers and /or professionals) referred to the adequacy of the items in order to measure a particular dimension of the tool (Crocker & Algina, 1986).

Scale developers often calculate the Osterlind index to provide empirical evidence of content validity, synthesizing the degree of agreement between experts about the fitness of each item to the dimension it measures (Osterlind, 1992). The classical procedure to calculate this index consists in the following steps:

Firstly, each expert gives a score on a 3-point rating scale to each item, being *-1* the lowest degree of suitability item-dimension, *0* the intermediate value, and +1 the highest degree of suitability.

Secondly, the information obtained from the evaluation of the different experts is operationalized using the Osterlind index of congruence (Rovinelli & Hambleton, 1977).

Its formal expression is:

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

Where:

X_{ijk} = Rating of item i in the dimension k by the judge j.

N = number of dimensions in the instrument.

n = number of judges.

An Osterlind index is calculated for each item. The possible results oscillate between ±*1*, depending on the degree of congruence in the expert answers. *-1* would implies that all the experts agree that the item does not fit to its dimension at all; *+1* would implies that all the experts assigned the highest degree of fitness item-dimension; *0* would be the lowest degree of agreement between expert opinions.

Finally, items which obtain *.5* or a higher score in the Osterlind index are usually included in the proposed test.

The aim of this work is to compare the results obtained when experts score the degree of suitability item-dimension on the traditional 3-point rating scales with the results obtained using, alternatively, 5-point ones as a revised formulation (Osterlind-R).

In this sense, we would test if one of the procedures is more conservative than the other

when taking the decision of considering the items adequate or not.

Method

Participants

105 students of *Psychometrics*, subject of the third course in the degree of Psychology at the University of Seville, participated in the expert judgment during the academic year 2009/10.

Instrument

The online questionnaire measured the construct «satisfaction with the training received». It was composed by 31 items distributed in seven dimensions. As an example, the first dimension was «aims/subject content», and 3 of their items were (a) *The clearness of the subject aims affects my satisfaction*; (b) *The difficulty of the subject contents affects my satisfaction*; and (c) *The interest of the subject contents affects my satisfaction.*

The written instructions presented were *Please, take a moment in order to complete this brief survey. The information you are providing is useful to improve the subject. Your answers are going to be treated confidentially. You have to judge the degree in which each item fits its dimension. Possible answers are as follows:(a) Strongly disagree; (b)Disagree;(c)Neither agree nor disagree;(d)Agree; and (e)Strongly agree.*

Procedure

In order to carry out the expert judgment, participants marked the degree each item fitted its dimension on a 5-point rating scale, and Osterlind-R index was calculated. Subsequently, the answers were transformed into a 3-point rating scale, and the classical Osterlind index was calculated. The number of possible options to choose (5 or 3) was the only difference between Osterlind-R and the classical Osterlind. The rest of the procedure (including the index calculation) did not differ. Table 1 presents the equivalence used to transform the scores:

Table 1

Note. Equivalence used to transform scores obtained from 5-point rating scales into 3-point rating scales

Osterlind-R criteria	Osterlind criteria
(-1) Strongly disagree	(-1) Disagree
(-.5) Disagree	
(0) Neither agree nor disagree	(0) Neither agree nor disagree
(+.5) Agree	(+1) Agree
(+1) Strongly agree	

Results

Table 2 presents the Osterlind index calculated for each item using the scores obtained on 5-point rating scales, and after transforming the scores into 3-point rating scales. All the values obtained in the first column (using 5-point rating scales) were lower than the obtained in the second column (using 3-point rating scales).

Table 2

Osterlind index using 5-point rating scales (Osterlind-R) and 3-point rating scales (Osterlind)

ITEM	OSTERLIND-R	OSTERLIND
1	.6810	.9238
2	.6333	.9238
3	.6905	.8571
4	.8143	.9143
5	.7286	.9333
6	.6095	.8857
7	.6952	.9333
8	.6714	.9238
9	.7667	.9143
10	.5476	.7714
11	.6810	.9238
12	.6143	.8667
13	.8190	.9619
14	.8048	.9524
15	.7333	.9143
16	.6381	.8476
17	.7619	.9238
18	.6429	.9048
19	.4286	.6746
20	.6952	.9048
21	.8190	.9333
22	.6714	.8381
23	.1810	.2381
24	.6810	.8190
25	.2619	.3714
26	.4476	.6571
27	.5667	.7810
28	.6333	.8762
29	.5476	.7905
30	.4429	.6095
31	.6524	.8857

Considering that .5 is the minimum value considered appropriate as congruence index Table 3 summarizes, using 5 and 3-point rating scales, the number of items that presented a lower value (from 0 to .5, excluding this last value), so they should be removed; and those that should be included as they presented a value of .5 or higher, distinguishing between the items that obtained values close to the limit (from .5 to .6 including the first value and excluding the last one), and those that clearly exceed the limit (from .6 to 1, including both values).

Table 3

Number of items using 5-point rating scales (Osterlind-R) and 3-point rating scales (Osterlind) removed and included considering .5 the minimum appropriate value

Index	Decission	Osterlind-R	Osterlind
[0-.5)	Removed	5	2
[.5-.6)	Included, close to be removed	3	0
[.6-1]	Clearly included	23	29

2 items (23 and 25) did not exceed the inclusion criterion (≥ 0.5) in both procedures (Osterlind-R and Osterlind); nevertheless, other 3 items (19, 26 and 30) were excluded using Osterlind-R, and included using the classical procedure.

3 items (10, 27 and 29) were located a bit over the inclusion criteria using the 5-point scale, while standing over .6 with the classical procedure.

There is a considerable increasing in the number of items over .6 (from 23 to 29 items) when using the 3-point scales.

Discussion

Using backward inference with a purposive sample of scale development studies (satisfaction questionnaire of the training received), we found that both methods could be used by researchers. We found considerable consistency item-dimension with both procedures. Nevertheless, the two approaches could lead to substantial different values, making it risky to draw conclusions about content validity.

Both procedures require a high level of agreement among experts, but one (Osterlind-R) is more conservative than the other; i.e., it is more difficult to find appropriate congruence indexes when using a 5-point rating scale.

These differences are enhanced regardless of the inclusion criterion used. With the classical .5, 26 items would be accepted with Osterlind-R and 29 with the classic procedure, while using for example .6, there would be 23 and 29 items accepted respectively.

Taking into account these results, it would be recommendable for scale developers to: (1) choose previously the number of points for the rating scales used (3 and 5), depending on their objectives and the level of restriction that they need when considering appropriate items; and (2) indicate the number of points presented in the rating scale used in order to provide to readers with interpretable content validity information.

If we assume that the values assigned to the possible answers given by experts to calculate the Osterlind index are carried out assuming a Gaussian function, we can consider that such behaviors can be modeled with the binomial distribution (for discrete variables) given the similarity between it and the Normal Law.

In this sense, the distribution of responses in both types of scales (5 and 3 points) would be the presented in Figures 1 and 2 respectively.

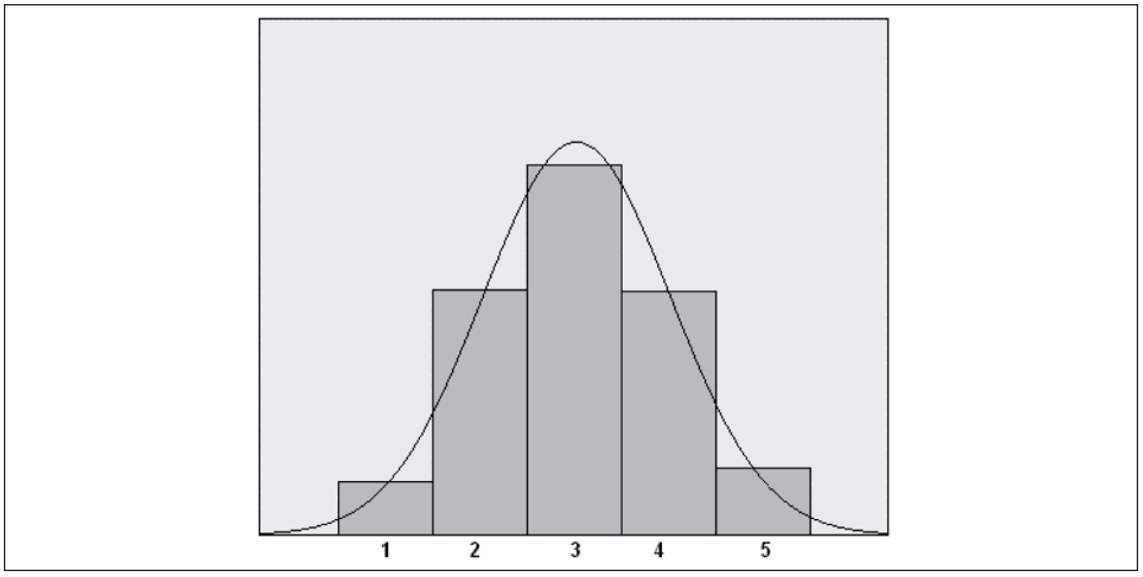

Figure 1. Distribution of responses in 5-point rating scales

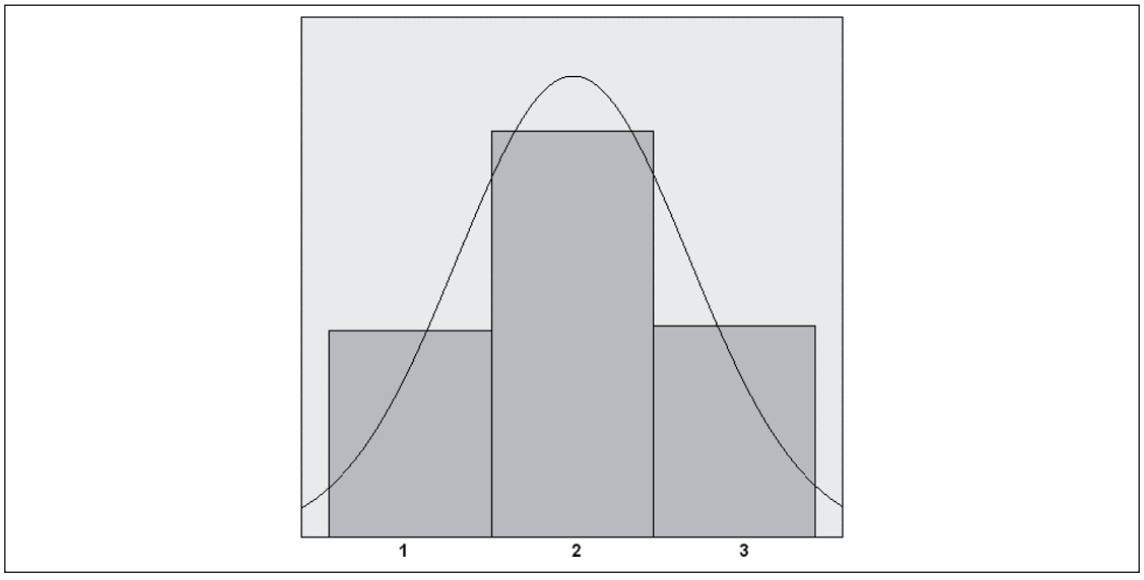

Figure 2. Distribution of responses in 3-point rating scales

Table 4 compares the different regions of probability when using a 5 or a 3-point rating escale.

Table 4

Regions of probability under the assumption of normality (P) when using a 5 or a 3-Note. Point rating scale.

Osterlind-R	P	Osterlind	P
(-1) Strongly disagree	.06	(-1) Disagree	.25
(-.5) Disagree	.25		
(0) Neither agree nor disagree	.38	(0) Neither agree nor disagree	.50
(+.5) Agree	.25	(+1) Agree	.25
(+1) Strongly agree	.06		

Note. Point rating scale.

The intermediate option «Neither agree nor disagree» presents a higher probability when 3-point rating scales are used (P =.5 vs. P =.38). As a consequence, the extreme options present more probability when 5-point rating scales are used; in Osterlind-R, adding the probability of the options «Strongly disagree» and «Disagree», or «Agree» and «Strongly agree», we obtain.06 +.25 =.31; while, when we use 3-option rating scales, the extreme options «Disagree» or «Agree» present a probability of.25. We consider that this is the key element that produces different values in the calculation of Osterlind index depending on the number of points that present the rating scales.

Some studies are planned for further research: (a) replications in different samples in order to increase the generalization of the results found; (b) the testing of different numbers of points in the rating scales (e. g., 4, 6 or 7) in order to study how they work; and (c) the gathering of data of the same sample answering based on 3 and 5-point rating scales in order to study the degree of influence of the number of points in the decisions taken by the participants, and how the possible differences affect the results obtained in the Osterlind index.

References

Anastasi, A. & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, NJ: Prentice Hall.

Carrasco, M. A., Holgado, F. P., Del Barrio, M. V., & Barbero, M. I. (2008). Incremental validity: an applied study using different informants and measurements. *Acción Psicológica, 5*, 65-76.

Chacón, S., Pérez-Gil, J. A., Holgado, F. P., & Lara, A. (1991). Evaluation of quality in higher education: Content validity. *Psicothema, 13*, 294-301.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Philadelphia: Harcourt Brace Jovanovich College Publishers.

Osterlind, S. J. (1992). *Constructing test items: multiple-choice, constructed-response, performance, and other formats*. Boston: Kluwer Academic Publishers.

Rovinelli, R. J. & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49-60.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.